طراحی فضایی برای انتخاب نقاط گره در مدل‌های دون‌رتبه

بهمن حمیدیان،[۱] حسین باغیشنی[۲]

تاریخ دریافت: ۱۳۹۵/۱۲/۲۳
تاریخ پذیرش: ۱۳۹۶/۱/۳۰

چکیده:

تحلیل بیزی داده‌های زمین‌آماری حجیم، با محاسبات ماتریسی سنگین و هزینه‌بر مواجه است. این محاسبات برای داده‌های فضایی و فضایی-زمانی چندمتغیره با ساختارهای وابستگی پیچیده، سنگین‌تر نیز خواهند بود. این مسئله برای الگوریتم‌های نمونه‌گیری MCMC که استفاده از آنها در تحلیل بیزی مدل‌های فضایی معمول هستند، مشکلاتی جدی مانند سرعت کند و همگرایی زنجیر ایجاد می‌کند. برای فرار از چنین مشکلات محاسباتی، یک رهیافت جانشین، استفاده از مدل‌های دون‌رتبه است که با کاهش فضای پارامتر و پرهیز از محاسبات ماتریسی سنگین، موجب می‌شود تا نرخ همگرایی الگوریتم‌های MCMC و سرعت محاسبات بهبود یابد. در مدل‌های دون‌رتبه، اطلاعات فضایی مکان‌های مشاهده‌شده در یک مجموعه از مکان‌های کوچک‌تر خلاصه می‌شوند. این مجموعهٔ کوچک‌تر به مجموعهٔ گره معروف است. تعیین نقاط مجموعهٔ گره و تعداد آنها به‌طوری که برآورد ساختار وابستگی فضایی متناظرشان نمایشی واضح و کم‌خطا از ساختار وابستگی حاصل از همهٔ داده‌ها باشد، یک جنبهٔ پایه‌ای و کلیدی در ساخت مدل‌های دون‌رتبه محسوب می‌شود. طراحی نقاط مکانی و تعداد گره‌ها برای اجرای این کاهش بعد، هدف اصلی این مقاله است. برای نمایش عملکرد طرح‌های مختلف در این رده از مدل‌ها، داده‌های کیفیت آب منطقهٔ وسیعی از استان گلستان را در بازهٔ زمانی سال‌های ۱۳۸۲ تا ۱۳۹۲ مورد تحلیل قرار داده‌ایم.

واژه‌های کلیدی: استنباط بیزی، الگوریتم MCMC، داده‌های فضایی-زمانی، مجموعهٔ گره، مدل دون‌رتبه.

۱ مقدمه

جامعهٔ علمی در حال حرکت به دورانی است که دسترسی به محیط‌های غنی از داده‌ها، فرصت‌های فوق‌العاده‌ای را فراهم می‌سازد تا به پیچیدگی‌ها و قابلیت‌های ساختارهای زمانی و مکانی فرایندها در مقیاس گسترده دست یابیم. گسترش داده‌های فضایی و فضایی-زمانی، موجب توسعهٔ قابل توجهی در مدل‌سازی‌های آماری شده است [۷و۸]. با توجه به ابزار مدرن گردآوری داده‌ها، از جمله سیستم‌های سنجش از راه دور، این نوع داده‌ها به‌طور فزاینده‌ای حجیم و متنوع هستند. نتایج به‌دست‌آمده از تحلیل این داده‌ها اغلب منجر به تصمیم‌گیری‌های مهمی در زمینه‌های مختلف اقتصادی، زیست‌محیطی، بهداشت و غیره می‌شوند. از طرفی، ساختارهای پیچیده و حجیم داده‌های حاصل از این فرایندها با مشکل انتخاب مدلی مناسب برای توصیف قابل قبول تغییرپذیری در سیستم‌های مورد علاقهٔ محققان مواجه هستند. اغلب مدل‌های موجود برای لحاظ کردن این پیچیدگی‌ها، ناقص هستند و توسعهٔ چارچوب مدل‌هایی که قادر به در نظر گرفتن منابع مختلف عدم قطعیت باشند نیز کار ساده‌ای نیست.

در حوزهٔ آمار فضایی، مدل‌های سلسله‌مراتبی [۸] کاربرد وسیعی در تحلیل داده‌های زمین‌آماری دارند. این مدل‌ها ساختار فضایی و زمانی داده‌ها را طی مراحل مختلف درون خود تعبیه می‌کنند و قادر هستند عدم قطعیت این نوع ساختارها را به‌خوبی لحاظ کنند. این مدل‌ها، به‌طور کلّی، از چارچوب استنباط بیزی پیروی می‌کنند[۱۶] و معمولاً تحلیل آنها بر اساس روش‌های نمونه‌گیری MCMC [۱و۲۶]، صورت می‌پذیرد که موجب محبوبیت این مدل‌ها در طیف وسیعی از کاربردها شده است. مشکل اصلی در تحلیل داده‌ها با مدل‌های سلسله‌مراتبی بیزی، سرعت کند الگوریتم‌های MCMC و تشخیص همگرایی زنجیر

[۱] دانش‌آموختهٔ کارشناسی ارشد آمار، دانشگاه صنعتی شاهرود، ایران
[۲] عضو هیئت علمی گروه آمار، دانشگاه صنعتی شاهرود، ایران

در حضور اندازهٔ بالای این نوع داده‌ها است. افزون بر این، وقتی بعد زمان اضافه و از مدل فضایی-زمانی استفاده شود، محاسبات پیچیده‌تر نیز می‌شوند. برای برازش یک مدل سلسله‌مراتبی با اندازهٔ نمونهٔ n، سهم محاسباتی اصلی، تجزیهٔ یک ماتریس همیشه‌مثبت با بعد n است. تعداد عملیات برای تجزیهٔ چنین ماتریسی از مرتبهٔ n^3 است و برای هر تکرار بیشتر، این تعداد عملیات به کل زمان اجرای الگوریتم افزوده می‌شود. به‌عنوان مثال فرض کنید برای یک اندازهٔ نمونهٔ مشخص، هر تکرار الگوریتم MCMC، ۰/۳ ثانیه زمان CPU را بگیرد. اگر ۱۰۰۰۰ تکرار برای همگرایی و استنباط کامل نیاز داشته باشیم، زمان کل مورد نیاز برای برازش مدل در حدود ۵۰ دقیقه خواهد بود. این زمان در مدل‌های پیچیدهٔ فضایی، با افزایش اندازهٔ نمونه، خیلی بیشتر خواهد شد. بنا بر این، برازش مستقیم مدل‌های دلخواه برای داده‌های فضایی و فضایی-زمانی حجیم، از نظر محاسباتی، ممکن نیست و باید از مدل‌های تقریبی استفاده کنیم.

رهیافت‌های مختلفی برای ارائهٔ مدل‌های تقریبی پیشنهاد شده‌اند که مبتنی بر پاسخ یک سؤال شکل گرفته‌اند: آیا رویه برازش‌شدهٔ حاصل از مدل تقریبی، یک کمیّت ثابت، مانند یک تابع پارامتری، است یا یک کمیّت تصادفی، یعنی تحققی از یک فرایند (میدان) تصادفی؟ در حالت اوّل، رویه به‌صورت نمایشی از مجموعه توابع پایه، مانند اسپلاین‌ها، بیان می‌شود. به‌عنوان مثال رویه را می‌توان به‌صورت

$$f(s) = \sum_{j=1}^{m} b_j g_j(s)$$

نوشت، که در آن b_jها، $j = 1, \ldots, m$، مجموعه‌ای از ضرایب و $\{g_j(s); j = 1, \ldots, m\}$ مجموعهٔ پایهٔ اسپلاین هستند. توابع اسپلاین بر حسب یک مجموعه از گره‌ها[3] تعریف می‌شوند که آنها را با $\{s_j^*, j = 1, \ldots, m \in D\}$ برچسب‌گذاری می‌کنیم. پژوهش‌های گسترده‌ای در این زمینه انجام شده‌اند. به‌عنوان چند نمونه می‌توان به [۴، ۵، ۱۷، ۲۱، ۲۵ و ۲۸] اشاره کرد. در این حالت، معمولاً یک گره، یکی از موقعیت‌های مشاهده‌شده در نظر گرفته می‌شود و کاهش بعدی صورت نمی‌گیرد.

در حالت دوم، رویه به‌صورت فرایند $\{\tilde{w}(s); s \in D\}$ تعریف می‌شود، به‌طوری که

$$\tilde{w}(s) = \sum_{j=1}^{m} l(s, s_j^*) Z(s_j^*).$$

در این تعریف، رویه (یا همان تحقق فرایند $\tilde{w}(\cdot)$) به‌طور کامل توسط تابع $l(\cdot, \cdot)$ و متغیرهای $\{Z(s_j^*)\ ,\ j = 1, \ldots, m\}$ تعیین می‌شود. در این حالت نیز s_j^*ها گره‌ها هستند و تعیین آنها (شامل تعداد و موقعیت آنها) یک مسالهٔ طراحی فضایی است. توجه داشته باشید که تعیین این گره‌ها هر دو تابع $l(\cdot, \cdot)$ و توزیع Zها را تحت تأثیر قرار می‌دهد. در این رده از مدل‌های تقریبی نیز رهیافت‌های مختلفی پیشنهاد شده‌اند. بدون ورود به جزئیات، می‌توان به [۱۲، ۱۳، ۲۰، ۳۰ و ۳۳] اشاره کرد. بسط و توسعه رهیافت‌های اشاره‌شده به مدل‌های فضایی سلسله‌مراتبی پیچیده مانند فرایندهای چندمتغیره [۳۴]، فضایی-زمانی [۱۵]، و مدل‌های با ساختارهای کوواریانسی ناایستا [۲۴]، دشوار است. در این موارد، استفاده از ردهٔ مدل‌های تقریبی دون‌رتبه مبتنی بر گره[4]، انتخابی معمول است [۲، ۹، ۱۸، ۲۲، ۳۱ و ۳۵].

ارائهٔ مدل‌های دون‌رتبه که حاصل روش‌های کاهش بعد می‌باشد، بر این ایده استوار است که اطلاعات فضایی مکان‌های مشاهده‌شده در یک مجموعه از مکان‌های کوچک‌تر قابل خلاصه شدن هستند؛ به این معنی که برای برازش مدل هنوز از همهٔ داده‌ها استفاده می‌کنیم، با این تفاوت که ساختار وابستگی فضایی (ماتریس کوواریانس) از طریق یک کاهش بعد نمایش داده می‌شود. این مجموعهٔ کوچک‌تر همان مجموعهٔ گره است. تعیین نقاط مجموعهٔ گره و تعداد آنها به‌طوری که برآورد ساختار وابستگی فضایی متناظرشان نمایشی واضح و کم‌خطا از ساختار وابستگی حاصل از همه داده‌ها باشد، یک جنبهٔ پایه‌ای و کلیدی در ساخت مدل‌های دون‌رتبه محسوب می‌شود. طراحی نقاط مکانی و تعداد گره‌ها برای اجرای این کاهش بعد، هدف اصلی این مقاله است.

در بخش ۲ ردهٔ مدل‌های دون‌رتبه و یک زیرردهٔ پرطرفدار از آن را با نام مدل‌های فرایند پیشگو[5] معرفی می‌کنیم. سپس در بخش ۳ روش‌های مرسوم و جدید تعیین نقاط گره (فضایی) را در مدل‌های فرایند پیشگو تشریح می‌کنیم. روش‌های معرفی‌شده در بخش ۳ را بر روی یک مجموعه دادهٔ فضایی-زمانی واقعی،

[3] knot set

[4] knot-based low-rank models

[5] predictive process models

مربوط به کیفیت آب‌های زیرزمینی استان گلستان، در بخش ۴ پیاده کرده، نتایج را ارزیابی و مقایسه می‌کنیم. در پایان نیز به نتیجه‌گیری مطالب مطرح‌شده می‌پردازیم.

۲ مدل‌های دون‌رتبه

در این بخش در ابتدا برخی از ویژگی‌های کاهش بعد را بیان می‌کنیم، سپس ردهٔ مدل‌های فرایند پیشگو را به‌عنوان یکی از زیررده‌های پرکاربرد مدل‌های دون‌رتبه معرفی می‌کنیم. میدان تصادفی

$$Z(\cdot)=\{Z(s);s\in D\}$$

را در نظر بگیرید که در آن $D\subset\Re^d$، $d\geq ۲$، ناحیهٔ جغرافیایی مورد نظر است. با پیروی از توضیحات بخش مقدمه، یک فرایند دون‌رتبه برای مجموعهٔ $\{s\in D\}$ به‌صورت

$$\tilde{w}(s)=\sum_{j=۱}^{m}l(s,s_j^*)Z(s_j^*) \qquad (۱)$$

تعریف می‌شود. قرار می‌دهیم $\tilde{\mathbf{w}}=(\tilde{w}(s_۱),\tilde{w}(s_۲),\ldots,\tilde{w}(s_n))^T$. در این صورت، فرایند (۱۱) به‌صورت زیر نوشته می‌شود:

$$\tilde{\mathbf{w}}=\mathbf{L}fz^* \qquad (۲)$$

که در آن $\mathbf{L}$ یک ماتریس $n\times m$ با درایهٔ (i,j)ام $l(s_i,s_j^*)$ و $\mathbf{z}^*$ برداری m-بعدی با عناصر $Z(s_j^*)$ است، به‌طوری که $m<<n$. با توجه به رابطهٔ (۱۲)، واضح است که به‌جای بردار n-بعدی $\tilde{\mathbf{w}}$، با بردار m-بعدی $\mathbf{z}^*$ سروکار داریم. بنا بر این با توجه به کوچک بودن m در مقابل n، کاهش بعد روشن است. افزون بر این، چون مدل را بر حسب $Z(s_j^*)$ها می‌نویسیم و با مفروض بودن $l(\cdot,\cdot)$ به‌طور قطعی $\tilde{w}$ها از روی Zها ساخته می‌شوند، ماتریس‌های کوواریانس متناظری که با آنها مواجه می‌شویم $m\times m$ هستند، نه $n\times n$. به‌طور واضح‌تر تابع کوواریانس معتبر به‌صورت

$$cov(\tilde{w}(s),\tilde{w}(s'))=\mathbf{l}(s)^T\Sigma_{\mathbf{z}^*}\mathbf{l}(s') \qquad (۳)$$

است، که در آن $\mathbf{l}(s)$ یک بردار m-بعدی با عناصر $l(s,s_j^*)$ می‌باشد. از (۱۳) مشاهده می‌کنید که حتی اگر تابع $l(\cdot,\cdot)$ ایستا باشد، یعنی بتوان آن را به‌صورت $l(\cdot-\cdot)$ نوشت، تابع کوواریانس القا شده برای $\tilde{w}(\cdot)$ ایستا نخواهد بود. همچنین اگر Zها گاوسی باشند، آن‌گاه $\tilde{w}(s)$ یک فرایند گاوسی خواهد بود.

اکنون چند سؤال اساسی حاضر می‌شوند: متغیرهای Z چه هستند؟ تابع $l(\cdot,\cdot)$ چگونه انتخاب می‌شود؟ و آیا گره‌های s_j^* یک زیرمجموعه از مکان‌های مشاهده‌شده s هستند یا به شکل دیگری انتخاب می‌شوند؟ سؤال سوم، ما را به مسالهٔ طراحی فضایی که موضوع اصلی این مقاله است، هدایت می‌کند و به آن در بخش ۳ پاسخ خواهیم داد. معمول‌ترین انتخاب برای Zها در نظر گرفتن آنها به‌عنوان متغیرهای i.i.d. از توزیع نرمال با میانگین صفر و واریانس $\sigma^۲$ است. به‌عبارت دیگر فرض می‌شود $Z(s)$ یک فرایند نوفهٔ سفید است. در این صورت، رابطهٔ (۷) به $\sigma^۲\mathbf{l}(s)^T\mathbf{l}(s')$ تبدیل می‌شود. از این انتخاب در[۳] و[۱۸] استفاده شده است. یک انتخاب طبیعی برای $l(\cdot,\cdot)$ نیز تابع هسته[۶] به‌شکل $K(s-s')$ است که بر اساس فاصلهٔ بین s و s' وزن‌دهی می‌کند. تابع هسته می‌تواند پارامتری باشد، که در نتیجه یک تابع کوواریانس پارامتری را القا می‌کند و همچنین می‌تواند به‌طور فضایی تغییرپذیر[۷] باشد [۱۸ و ۲۴].

با در نظر گرفتن تابع هسته برای l، شکل فرایند دون‌رتبه (۱۱) را می‌توان به‌عنوان تقریبی گسسته از نمایش فرایند به‌شکل $\tilde{w}(s)=\int_{R^d}K(s-s')Z(s')ds'$ تصور کرد [۳۶]. توابع هستهٔ گاوسی که تابع کوواریانس گاوسی را نتیجه می‌دهند، بیشتر مورد استفاده قرار می‌گیرند [۲۹]. یکی از معایب جدی تعریف فرایندهای دون‌رتبه بر اساس تابع هسته، محدودیت کاربردی آنهاست. برای نمونه، تابع کوواریانس نمایی که به‌طور گسترده‌ای برای تحلیل داده‌های فضایی مورد استفاده قرار می‌گیرد، با این روش قابل دست‌یابی نیست.

یک رهیافت متفاوت برای تعیین Z، در نظر گرفتن آن از یک فرایند تصادفی با یک تابع کوواریانس مشخص است. با این انتخاب، بنا به (۱۳)، یک تابع کوواریانس جدید به $\tilde{w}(\cdot)$ القا می‌شود. اکنون این سؤال مطرح می‌شود که اگر علاقه‌مند باشیم $\tilde{w}(s)$ یک تابع کوواریانس دلخواه مشخص داشته باشد، باید چه تابع کوواریانسی را برای Zها انتخاب کنیم؟ برای ارائهٔ پاسخی بهینه به این سؤال، در ادامه مدل‌های فرایند پیشگو (گاوسی) را معرفی می‌کنیم که توسط [۲] معرفی شد.

[۶] kernel function

[۷] spatially varying

۱.۲ مدل‌های فرایند پیشگو

مدل رگرسیونی خطی برای پاسخ $y(\mathbf{s}_i)$، $\mathbf{s}_i \in D$، را به‌صورت

$$y(\mathbf{s}_i) = \mathbf{x}(\mathbf{s}_i)^T\beta + w(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i),\ i = ۱,\dots,n \qquad (۴)$$

در نظر بگیرید، که در آن $\mathbf{x}(\mathbf{s}_i) = (x_۱(\mathbf{s}_i),\dots,x_p(\mathbf{s}_i))^T$ بردار متغیرهای تبیینی p-بعدی، $\beta = (\beta_۱,\dots,\beta_p)^T$ بردار ضرایب رگرسیونی (اثرهای ثابت) p-بعدی و $w(\mathbf{s})$ اثر تصادفی فضایی است. مولفهٔ $\varepsilon(\mathbf{s})$ نیز فرایند خطای اندازه‌گیری است که فرض می‌کنیم عناصر آن برای $i = ۱,\dots,n$، مستقل و هم‌توزیع با توزیع نرمال $N(۰,\tau^۲)$ هستند. معمولاً $w(\mathbf{s})$ یک فرایند گاوسی با تابع میانگین ثابت ۰ و تابع کوواریانس $C(\mathbf{s},\mathbf{s}')$ انتخاب می‌شود، که از این به بعد این فرایند گاوسی را با نماد $GP\{۰, C(\mathbf{s},\mathbf{s}')\}$ نمایش می‌دهیم. در عمل، اغلب تابع کوواریانس به‌صورت پارامتری $C(\mathbf{s},\mathbf{s}';\theta) = \sigma^۲\rho(\mathbf{s},\mathbf{s}';\theta^*)$ در نظر گرفته می‌شود، که در آن $\rho(\cdot,\cdot;\theta^*)$ تابع همبستگی‌نگاشت و θ^* بردار q-بعدی پارامترهای وابستگی فرایند فضایی است و همچنین $\theta = (\sigma^۲,\theta^*)^T$.

مجموعهٔ گره $S^* = \{\mathbf{s}^*_۱,\dots,\mathbf{s}^*_m\}$ را در نظر بگیرید که ممکن است یک زیرمجموعه از مکان‌های مشاهده‌شده $S = \{\mathbf{s}_۱,\dots,\mathbf{s}_n\}$ باشد. تعریف می‌کنیم

$$\mathbf{w}^* = (w(\mathbf{s}^*_۱),\dots,w(\mathbf{s}^*_m)) \sim GP\{۰, C^*(\theta)\}\ ,$$

که در آن $C^*(\theta) = \big(C(\mathbf{s}^*_i,\mathbf{s}^*_j;\theta)\big)$، $i,j = ۱,\dots,m$، ماتریس کوواریانس $m \times m$ متناظر با مجموعهٔ گره است. چون فرایند تعریف‌شده در (۱۴) گاوسی فرض شد، فرایند $\mathbf{w}^*$ نیز یک فرایند گاوسی خواهد بود. پیشگوی فضایی در مکان دلخواه $\mathbf{s}.$ با

$$\tilde{w}(\mathbf{s}.) = E\left[w(\mathbf{s}.)|\mathbf{w}^*\right]$$

تعریف می‌شود. از آن‌جا که توزیع توأم $w(\mathbf{s}.)|\mathbf{w}^*$ نرمال است، به سادگی نتیجه می‌شود

$$\tilde{w}(\mathbf{s}.) = E\left[w(\mathbf{s}.)|\mathbf{w}^*\right] = c^T(\mathbf{s}.;\theta)C^{*-۱}(\theta)\mathbf{w}^*,$$

که در آن $c^T(\mathbf{s}.;\theta) = (C(\mathbf{s}.,\mathbf{s}^*_۱;\theta),\dots,C(\mathbf{s}.,\mathbf{s}^*_m;\theta))$ است. فرایند $\tilde{w}(\mathbf{s})$، را فرایند پیشگوی به‌دست‌آمده از فرایند والد[۸] $w(\mathbf{s})$ می‌نامند. بنا بر این

$$\tilde{w}(\mathbf{s}) \sim GP\left\{۰, \tilde{C}(\theta)\right\}$$

که در آن

$$\begin{aligned}\tilde{C}(\theta) &= \tilde{C}(\mathbf{s},\mathbf{s}';\theta) = Cov(\tilde{w}(\mathbf{s}),\tilde{w}(\mathbf{s}'))\\ &= Cov(c^T(\mathbf{s};\theta)C^{*-۱}(\theta)\mathbf{w}^*, c^T(\mathbf{s}';\theta)C^{*-۱}(\theta)\mathbf{w}^*)\\ &= c^T(\mathbf{s};\theta)C^{*-۱}(\theta)Var(\mathbf{w}^*)C^{*-۱}(\theta)c(\mathbf{s}';\theta)\\ &= c^T(\mathbf{s};\theta)C^{*-۱}(\theta)C^*(\theta)C^{*-۱}(\theta)c(\mathbf{s}';\theta)\\ &= c^T(\mathbf{s};\theta)C^{*-۱}(\theta)c(\mathbf{s}';\theta). \qquad (۵)\end{aligned}$$

فرایند پیشگو توسط تابع کوواریانس فرایند والد و مجموعهٔ گره S^* تعیین می‌شود. پس در واقع باید بنویسیم $\tilde{w}_{S^*}(\mathbf{s})$، ولی آن را به‌اختصار با $\tilde{w}(\mathbf{s})$ نمایش می‌دهیم. همچنین با توجه به تابع کوواریانس (۵)، فرایند پیشگو، صرف نظر از ایستا بودن یا نبودن $w(\mathbf{s})$، ایستا نیست.

با جایگزینی $w(\mathbf{s})$ با $\tilde{w}(\mathbf{s})$ در مدل (۱۴)، مدل فرایند پیشگو، به‌عنوان یک مدل دون‌رتبه، به‌صورت زیر نتیجه می‌شود:

$$y(\mathbf{s}_i) = \mathbf{x}(\mathbf{s}_i)^T\beta + \tilde{w}(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i). \qquad (۶)$$

از آن‌جا که

$$\tilde{w}(\mathbf{s}) = c^T(\mathbf{s};\theta)C^{*-۱}(\theta)\mathbf{w}^*$$

بنا بر این فرایند $\tilde{w}(\mathbf{s})$ یک تبدیل خطی فضایی از $\mathbf{w}^*$ است. در برازش مدل (۶)، n اثر تصادفی $\{w(\mathbf{s}_i), i = ۱,\dots n\}$ با m اثر تصادفی $\mathbf{w}^*$ جایگزین شده‌اند و ما با یک توزیع توأم نرمال m بعدی شامل یک ماتریس کوواریانس $m \times m$ سر و کار داریم. بنا بر این کاهش بعد به‌روشنی مشاهده می‌شود. این کاهش بعد باعث کاهش هزینه‌های محاسباتی و در نتیجه کارایی محاسباتی می‌شود. در مقابل بخشی از کارایی آماری مدل را از دست می‌دهیم. توجه کنید که اگرچه از پارامترهای مشابه در هر دو مدل استفاده می‌کنیم، اما در واقع پارامترهای دو مدل با هم یکسان نیستند و مدل (۶) با مدل (۱۴) متفاوت است.

۳ طراحی نقاط گره

در این بخش، مروری بر روش‌های انتخاب گره همراه با روشی که بر پایه مقالهٔ [۱۰] طراحی شده است، ارائه می‌کنیم.

[۸] parent process

۱.۳ مروری کوتاه بر طرح‌های نمونه‌گیری فضایی

مشابه هر روش مبتنی بر گره، انتخاب گره‌ها در فرایندهای پیشگو یک مسالهٔ چالش‌برانگیز است. این مسئله به این دلیل که گره‌ها d-بعدی (و معمولاً دوبعدی) هستند، نسبت به حالت تک بعدی در روش‌هایی مثل هموارسازی اسپلاین، مشکل‌تر است.

در ابتدا فرض کنید m معلوم است. در متون مربوط به‌روش‌های هموارسازی اسپلاین (و در اکثر متون مربوط به داده‌های تابعی یا مدل‌بندی رگرسیونی با استفاده از نمایش توابع پایه)، انتخاب مشاهدات به‌عنوان گره‌ها مرسوم است [۲۵]. اما این انتخاب در مدل فرایند پیشگو به‌عنوان یک گزینه، قابل دفاع نیست و این سؤال مطرح می‌شود که آیا از یک زیرمجموعه از مکان‌های فضایی مشاهده‌شده استفاده کنیم یا یک مجموعهٔ جدا از مکان‌های مشاهده‌شده یا ترکیبی از هر دو؟

- اگر یک زیرمجموعه از مکان‌های نمونه‌گیری‌شده انتخاب کنیم، آیا این انتخاب باید تصادفی باشد؟
- اگر نخواهیم از زیرمجموعه‌ای از مکان‌های مشاهده‌شده استفاده کنیم، یک مسالهٔ طرح نمونه‌گیری فضایی[۹] در حضور n نقطه نمونه‌گیری‌شده پیش می‌آید. البته در زمینه طراحی نمونه‌گیری فضایی، تحقیقات گسترده‌ای صورت گرفته که در [۳۶] خلاصه شده‌اند.

یک روش انتخاب گره که پرکردن فضا[۱۰] نام دارد، توسط [۲۳] معرفی شد. این طرح از جمله طرح‌های تعریف‌شده بر پایهٔ معیارهای هندسی است. انگیزهٔ تعریف این معیارها بر این مبنا است که توانایی پوشش ناحیهٔ مورد مطالعه توسط یک مجموعه از نقاط انتخابی، مستقل از تابع کوواریانس فضایی، چه‌قدر است؟ در حالت خاص فرض می‌شود پراکنش نقاط بر روی ناحیهٔ D از توزیع یکنواخت پیروی می‌کند و m نقطه از این توزیع انتخاب و از آنها به‌عنوان مجموعهٔ گره استفاده می‌شود.

در طراحی نمونه‌گیری فضایی و فضایی-زمانی، منابع غنی وجود دارند. [۳۶] الگوریتم‌هایی مانند انتخاب دنباله‌ای، انتخاب بلوکی و جستجوی تصادفی را پیشنهاد کردند. [۳۸] طرح‌های بهینه‌ای بر اساس تابع درست‌نمایی معرفی کردند. آنها اطلاع فیشر را به‌عنوان یک معیار انتخاب طرح بهینه پیشنهاد کردند. [۱۰] نیز طرح‌های مختلف نمونه‌گیری فضایی را مورد کاوش قرار دادند و استفاده از طرح‌های اصلاح‌شدهٔ مشبکهٔ منظم را توصیه کردند. در ادامه، برخی از این روش‌ها را توضیح خواهیم داد. یک طرح نمونه‌گیری مشبکهٔ منظم، یک ماتریس $k \times f$ را شامل می‌شود که در آن گره‌ها با فاصلهٔ یکسان از یکدیگر قرار گرفته‌اند. دو ردهٔ اصلاح‌شده از این نوع طرح‌ها، مشبکهٔ منظم همراه با جفت‌های نزدیک[۱۱] و مشبکهٔ منظم طرح پر[۱۲] نام دارند.

۱.۱.۳ طرح نمونه‌گیری مشبکهٔ منظم همراه با جفت‌های نزدیک

یک مشبکهٔ منظم $k \times k$ را در نظر بگیرید. فرض کنید نقاط مشبکه با فاصلهٔ Δ از یکدیگر قرار داشته باشند. مشبکه را با انتخاب تصادفی m' نقطه و قرار دادن یک نقطهٔ اضافی نزدیک به هر یک از این نقاط انتخابی، تقویت می‌کنیم. هر کدام از نقاط یا مکان‌های جدید ایجاد شده، از یک دایره به مرکز m' و شعاع $\delta = \alpha\Delta$ به‌طور تصادفی انتخاب می‌شوند. ما برای نمایش آن از نماد $(k \times k, m', \alpha)$ استفاده می‌کنیم و انتخاب مقدار α اختیاری است.

۲.۱.۳ مشبکهٔ منظم طرح پر

یک شبکهٔ منظم $k \times k$ را در نظر بگیرید. مثل روش قبل، m' نقطه از مشبکه را به‌طور تصادفی انتخاب می‌کنیم و در نقاط انتخابی یک مشبکهٔ کوچک‌تر با فواصل نزدیک‌تر به هم طراحی می‌کنیم. برای نمایش آن از نماد $(k \times k, m', r \times r)$ استفاده می‌کنیم که انتخاب مقدار r اختیاری است. هر مشبکهٔ جدید ایجاد شده یک مشبکهٔ $r \times r$ و شامل $r^2 - ۴$ مکان اضافی است.

۲.۳ راهکاری جانشین برای انتخاب گره

[۱۱] روشی را برای انتخاب گره پیشنهاد کردند که هدف آن بهبود فرایند پیشگوی القا شده به‌عنوان یک تقریب از فرایند والد است. برای یک مجموعهٔ گره، فرایند $\tilde{w}(\mathbf{s}) = E[w(\mathbf{s})|\mathbf{w}^*]$ را

[۹] spatial design

[۱۰] space-filling

[۱۱] lattice plus close pairs

[۱۲] lattice plus infill

به‌عنوان یک تقریب از فرایند والد در نظر بگیرید. با شرط مفروض بودن بردار پارامترهای θ، واریانس این فرایند پیشگو برابر است با

$$Var_\theta(s,S^*) = Var(w(\mathbf{s})|\mathbf{w}^*,\theta) = C(\mathbf{s},\mathbf{s};\theta) - c^T(\mathbf{s};\theta)C^{*-1}(\theta)c(\mathbf{s};\theta). \quad (۷)$$

این واریانس شرطی معیاری برای اندازه‌گیری قدرت تقریب فرایند والد توسط فرایند پیشگو است. هر قدر واریانس شرطی کمتر باشد، مجموعهٔ گره انتخابی بهتر خواهد بود. در حالت‌های خاص، معیار انتخاب گره به‌عنوان تابعی از $Var_\theta(s,S^*)$ تعریف می‌شود. یک معیار معمول به‌صورت

$$Var_\theta(S^*) = \int_D Var_\theta(s,S^*)g(s)ds = \int_D Var(w(\mathbf{s})|\mathbf{w}^*,\theta)g(s)ds \quad (۸)$$

تعریف می‌شود، که در آن $g(\cdot)$ یک تابع انتگرال‌پذیر روی D است و $g(s)$ وزن اختصاص‌یافته به مکان s است [۱۰]. در این مقاله حالت سادهٔ $g(s)=1$ را در نظر می‌گیریم. بنا بر این، معیار مورد نظر، میانگین واریانس پیشگوی فضایی خواهد شد. محاسبهٔ تحلیلی انتگرال (۱۶) ممکن نیست و باید با روش‌های عددی مانند تربیع‌بندی عددی یا انتگرال‌گیری مونته‌کارلویی تقریب زده شود. شکل تقریب مورد استفادهٔ ما به‌صورت

$$Var_\theta(S^*) \approx \frac{\sum_{i=1}^{n} Var(w(\mathbf{s})|\mathbf{w}^*,\theta)}{n} \quad (۹)$$

است. بنا بر این، مجموعهٔ گرهی انتخاب می‌شود که معیار طراحی $Var_\theta(S^*)$ را کمینه کند.

در مدل فرایند والد، ماتریس کوواریانس $\Sigma_Y = C(\theta)+\tau^2 I$ است، در حالی‌که ماتریس متناظر در فرایند پیشگوی حاصل از فرایند والد $\tilde{\Sigma}_Y = c^T(\mathbf{s};\theta)C^{*-1}(\theta)c(\mathbf{s};\theta)+\tau^2 I$ می‌باشد. نرم فروبنیوس[۱۳] بین Σ_Y و $\tilde{\Sigma}_Y$ برابر است با

$$||\Sigma_Y - \tilde{\Sigma}_Y||_F \equiv tr\left(\left[C(\theta) - c^T(\mathbf{s};\theta)C^{*-1}(\theta)c(\mathbf{s};\theta)\right]\right)^2.$$

از آن‌جایی که عبارت زیر توان دوم در داخل تابع اثر، یک ماتریس همیشه‌مثبت است، داریم $||\Sigma_Y - \tilde{\Sigma}_Y||_F = \sum \lambda_i^2$ که λ_iها مقادیر ویژهٔ ماتریس $\Sigma_Y - \tilde{\Sigma}_Y$ هستند. در نهایت، میانگین واریانس پیشگوی فضایی برابر است با

$$\bar{V} = \frac{1}{n} tr(\Sigma_Y - \tilde{\Sigma}_Y) = \frac{1}{n}\sum \lambda_i^2.$$

توجه کنید که حتی صورت تقریبی $Var_\theta(S^*)$ را نمی‌توان محاسبه کرد، زیرا به پارامتر نامعلوم θ وابسته است. یک راه برای رفع این مشکل، برآورد θ توسط زیرمجموعه‌ای از داده‌های مشاهده‌شده است. راه دیگر، استفاده از رهیافت بیزی است به‌طوری که یک پیشین برای θ تعریف کنیم و سپس معیار $E_\theta(Var_\theta(S^*))$ را مینیمم کنیم [۱۰].

حال فرض کنید مقادیر پارامترها و تعداد گره‌ها، m، مشخص باشند. الگوریتم جستجوی دنباله‌ای زیر، تقریبا یک طرح نمونه‌گیری بهینه را نتیجه می‌دهد:

۱. مجموعه نمونهٔ مجاز: فرض کنید ناحیهٔ مورد مطالعه پیوسته باشد. برای اجرای طرح نمونه‌گیری بهینه باید مکان‌های نمونه‌گیری ممکن را به یک مجموعهٔ متناهی کاهش دهیم. انتخاب‌های معمول شامل مجموعهٔ گره مشبکه، مجموعه مکان‌های مشاهده‌شده یا اجتماع این دو مجموعه می‌باشند.

۲. مجموعهٔ گره اوّلیه: یک مجموعه مکان اوّلیه را با اندازه $m_0 \leq m$ به‌عنوان نقاط آغازین برای انتخاب گره مشخص می‌کنیم. انتخاب ممکن می‌تواند یک مشبکه یا زیرمجموعه‌ای از مکان‌های مشاهده‌شده به‌صورت تصادفی یا قطعی باشد.

۳. در مرحله $t+1$

- برای هر نمونهٔ s_i در مجموعه نمونهٔ مجاز، $Var_\theta(s_i,S^{*(t)})$ را حساب می‌کنیم.
- نمونه‌ای را که بیشترین کاهش در $\bar{V}$ را دارد از مجموعه نمونهٔ مجاز حذف و به مجموعهٔ گره اضافه می‌کنیم.

۴. مراحل بالا را تا رسیدن به m گره تکرار می‌کنیم.

شایان ذکر است که این الگوریتم، یک طراحی بهینه است، ولی یک راه‌حل دائمی (عمومی) بهینه نیست. راه‌حل‌های دیگری نیز برای تقریب طراحی نمونه‌گیری بهینه مانند جستجوی تصادفی و انتخاب بلوکی در دسترس هستند [۳۶].

در نحوهٔ انتخاب تعداد گره یعنی m، پاسخ «تا حد ممکن بزرگ» را داریم. اما واضح است که انتخاب m به هزینهٔ محاسباتی و حساسیت استنباط بستگی دارد. بنا بر این، در عمل باید تحلیل را با mهای مختلف انجام دهیم و نتایج را با هم، بر اساس معیارهای تعریف‌شده، مقایسه کنیم. مقایسه باید نسبی

[۱۳] frobenius norm

باشد و زمان اجرای محاسبات در فرایند استنباط، در نظر گرفته شود. در کنار این‌ها می‌توان کاهش $\bar{V}$ با mهای مختلف را نیز به‌عنوان یک معیار در نظر گرفت.

۳.۳ معیارهای سنجش طرح‌های نمونه‌گیری

برای سنجش بهینگی طرح‌های نمونه‌گیری فضایی، معیارهای مختلفی معرفی شده‌اند. سه معیار پرکاربرد انتخاب مدل بیزی با نام‌های G، P و D توسط [۱۴] معرفی شدند که برای انتخاب طرح نمونه‌گیری بهتر در این مقاله از آنها بهره می‌بریم. برای محاسبهٔ این معیارها از مدل برازش‌شده توسط داده‌های طرح کمک گرفته می‌شود. برای تشریح مسئله، مدل (۶) را در نظر بگیرید. برای هر $s_i \in S$، از مدل

$$N\left(x(s_i)^T\beta^{(\ell)} + \tilde{w}(s_i)^{(\ell)}, \tau^{۲(\ell)}\right), \qquad \ell = ۱, \ldots, B$$

مشاهدات مستقل $y_{rep}(s_i)^{(\ell)}$ تولید می‌شوند، که در آن $(\beta^{(\ell)}, \tilde{w}(s_i)^{(\ell)}, \tau^{۲(\ell)})$، $\ell = ۱, \ldots, B$، نمونه‌های پسین تولید شده هستند. با قرار دادن $\mu_{rep,i}$ و $\sigma^۲_{rep,i}$ به‌عنوان میانگین و واریانس حاصل از $y_{rep}(s_i)$ ها، معیارهای مذکور به‌صورت زیر تعریف می‌شوند:

$$\begin{aligned} G &= \sum_{i=۱}^{n} \left(y(s_i) - \mu_{rep,i}\right)^۲, \\ P &= \sum_{i=۱}^{n} \sigma^۲_{rep,i}, \\ D &= G + P. \end{aligned}$$

مقادیر کوچک‌تر این معیارها، بیان‌کنندهٔ طرح‌های نمونه‌گیری بهتر هستند.

۴ کیفیت آب‌های زیرزمینی استان گلستان

توانایی اندازه‌گیری آب برای عبور جریان الکتریکی، هدایت الکتریکی[۱۴] (EC) نامیده می‌شود. هدایت الکتریکی، تابعی از غلظت آنیون‌ها، کاتیون‌ها و دمای محلول بوده، با مقدار کل جامدات محلول[۱۵] (TDS) همبستگی دارد و یکی از سریع‌ترین روش‌های ارزیابی کلّی کیفیت آب‌های زیرزمینی است. قابلیت هدایت الکتریکی یا عکس آن، یعنی مقاومت الکتریکی، یکی از عوامل مهم در تحقیقات هیدروژئولوژی است و در اندازه‌گیری آن احتیاجی به دستگاه‌های پچیده نیست. بنا بر این مشکلات نمونه‌برداری وجود ندارد و اندازه‌گیری آن سریع و دقیق است. مقدار استاندارد تعیین‌شده برای EC توسط سازمان جهانی بهداشت، ۵۰۰ میکروموس است. با توجه به این ویژگی‌ها، هدف این مثال موردی، ارزیابی طرح‌های مختلف نمونه‌گیری نقاط گره در مدل‌های فرایند پیشگو برای تحلیل کیفیت آب‌های زیرزمینی بخشی از استان گلستان بر پایهٔ متغیر پاسخ EC است. داده‌های موجود، شامل اندازه‌های ثبت‌شده در ۱۵۰ ایستگاه از آب‌های زیرزمینی استان گلستان است که از آبان ۱۳۸۲ تا آبان ۱۳۹۲ گردآوری شده‌اند. منظور از ایستگاه‌های آب زیرزمینی، چشمه‌ها و چاه‌های عمیق و نیمه‌عمیق است و نمونه‌گیری نیز در دو ماه از سال، یعنی اردیبهشت و آبان انجام گرفته است. بنا بر این مجموعه‌داده‌ها شامل ۱۵۰ نقطهٔ مکانی و ۲۱ نقطهٔ زمانی و در مجموع ۳۱۵۰ مشاهدهٔ فضایی-زمانی برای هر متغیر است. متغیر پاسخ EC و متغیرهای تبیینی شامل سختی کل[۱۶] (TH)، پتاسیم (K)، سدیم (Na)، سولفات ($SO_۴$)، کلر (Cl)، بیکربنات ($HCO_۳$)، ph و TDS هستند.

بر اساس تحلیل اکتشافی داده‌ها (که به‌دلیل محدودیت اندازهٔ مقاله گزارش نشده است)، تابع کوواریانس نمایی با برازشی خوب برای ساختار وابستگی فضایی انتخاب شد.

برای تحلیل این داده‌ها از یک مدل خطی فضایی-زمانی پویا استفاده کردیم که حالت خاصی از مدل‌های فضایی-زمانی است. فرض کنید $y_t(s)$ متغیر پاسخ در مکان s و زمان t باشد، به‌طوری که $t = ۱, ۲, \ldots, N_t$ و $s \in D \subseteq R^۲$. مدل مورد نظر ما در قالب یک مدل فضای حالتی، ترکیبی از دو دسته معادلهٔ مشاهده و معادله‌های حالت است که از آنها برای وارد کردن مؤلفه‌های رگرسیونی و فضایی-زمانی به مدل، استفاده و به‌صورت زیر نوشته

[۱۴] electric conductivity

[۱۵] total dissolved solution

[۱۶] total hardness

می‌شوند:

$$y_t(s) = x_t(s)^T\beta_t + u_t(s) + \varepsilon_t(s), \qquad \varepsilon_t(s) \overset{ind.}{\sim} N(۰, \tau^۲);$$
$$\beta_t = \beta_{t-۱} + \eta_t, \qquad \eta_t \overset{i.i.d.}{\sim} N(۰, \Sigma_\eta);$$
$$u_t(s) = u_{t-۱}(s) + w_t(s), \qquad w_t(s) \overset{ind.}{\sim} GP(۰, C_t(\cdot, \theta_t)) \quad (۱۰)$$

برای $t = ۱, ۲, \ldots, N_t$، که اختصارات .ind و .i.i.d به‌ترتیب نشان‌دهندهٔ «مستقل» و «مستقل و هم‌توزیع» هستند. در این‌جا $x_t(s)$ بردار p-بعدی متغیرهای تبیینی و β_t بردار p-بعدی ضرایب رگرسیونی وابسته به زمان متناظر با متغیرهای تبیینی می‌باشد. نماد $GP(۰, C_t(\cdot, \theta_t))$ نیز نمایانگر فرایند گاوسی با تابع کوواریانس پارامتری $C_t(\cdot; \theta_t)$ است. معمولاً شکل تابع کوواریانس به‌صورت $C_t(s_۱, s_۲; \theta_t) = \sigma_t^۲\rho(s_۱, s_۲; \phi_t)$ است، که در آن $\theta_t = \{\sigma_t^۲, \phi_t\}$، $\rho(\cdot; \phi_t)$ تابع همبستگی با پارامتر ϕ_t، و $\sigma_t^۲$ مولفهٔ واریانس فضایی می‌باشد. یک انتخاب برای تعریف تابع همبستگی فضایی، تابع نمایی است که به‌صورت $C_t(s_۱, s_۲; \theta_t) = \sigma_t^۲ \exp\{-\phi_t||s_۱ - s_۲||\}$ نمایش داده می‌شود و در آن $||s_۱ - s_۲||$ فاصلهٔ اقلیدسی بین مکان‌های $s_۱$ و $s_۲$ است. با این حال، هر تابع همبستگی فضایی معتبر را نیز می‌توان به کار برد. با پذیرفتن $\beta_. \sim N(m_., \Sigma_.)$ و $u_.(s) = ۰$، برای $t = ۰$، مدل سلسله‌مراتبی بیزی کامل می‌شود.

حال برای اجرای تحلیل بیزی، باید برای ضرایب مدل، توزیع پیشینی انتخاب کنیم. یک انتخاب معمول توزیع پیشینی برای ضرایب رگرسیونی، وابستگی فضایی و Σ_η به‌ترتیب توزیع نرمال، گامای معکوس و ویشارت معکوس می‌باشد و ما فرض کردیم که در این مثال، توزیع‌های پیشین از این توزیع‌ها و با پارامترهای زیر آمده‌اند و سپس به تحلیل مدل پرداختیم.

برای ضرایب رگرسیونی، توزیع پیشینی نرمال $\sim \beta$ $N(۰, I_۸)$ را انتخاب کردیم که در آن $I_۸$ ماتریس همانی با بعد ۸ است. همچنین برای ضرایب وابستگی فضایی توزیع گامای معکوس $\phi_t, \sigma_t^۲, \tau_t^۲ \sim IG(۲, ۲۵)$ انتخاب شد. علاوه بر این فرض کردیم $\Sigma_\eta \sim IW(۲, diag(۰/۰۰۱, ۸))$ که در آن $diag(۰/۰۰۱, ۸)$ یک ماتریس قطری با بعد ۸ با درایه‌های روی قطر ۰/۰۰۱ است. تعداد ۲۵ و ۵۰ نقطهٔ مکانی به‌عنوان نقاط گره مطابق با روش‌های مشبکهٔ منظم، طرح نمونه‌گیری پرکردن فضا و تقریب طراحی بهینه، انتخاب شدند که در شکل ۱ نمایش داده شده‌اند. سپس مدل فرایند پیشگو را با این مشخصات برازش دادیم. همان‌طور که در شکل ۱ دیده می‌شود هر سه روش انتخاب گره به‌خوبی توانسته‌اند، در مدل فرایند پیشگو، رویهٔ واقعی مدل را بازیابی کنند. همچنین نتایج مقایسهٔ سه روش طراحی نمونه‌گیری نقاط گره (با تعداد گره‌های متفاوت) بر اساس معیارهای ذکر شده و همچنین میانگین توان دوم خطاهای پیشگویی (MSPE) و زمان لازم برای برازش مدل (به ثانیه)، در جدول ۱ گزارش شده‌اند. با توجه به نتایج به‌دست‌آمده می‌توان موارد زیر را بیان کرد:

- علاوه بر به دست آوردن کارایی محاسباتی قابل توجه در قیاس با مدل رتبهٔ کامل (کاهش زمانی قابل توجه نسبت به مدل رتبهٔ کامل) کارایی آماری بسیار خوبی از نتایج استنباط مشهود است.
- با افزایش تعداد نقاط گره، دقت پیشگویی افزایش یافته است. البته با دو برابر شدن نقاط گره، زمان اجرای الگوریتم برازش مدل سه برابر شده است.
- به‌طور کلّی، عملکرد هر سه روش با تعداد نقاط گره یکسان بر اساس همه معیارهای مقایسه، مشابه است.

جدول ۱. مقایسهٔ مدل‌های فرایند پیشگو با طراحی و تعداد گره‌های مختلف

مدل رتبهٔ کامل	مشبکهٔ منظم		پرکردن فضا		طراحی بهینه	
	۵۰ گره	۲۵ گره	۵۰ گره	۲۵ گره	۵۰ گره	۲۵ گره
۳۰/۴۱	۲۴/۴۷	۲۴/۹۸	۲۴/۴۷	۲۵/۷۹	۲۴/۹	۲۵/۰۵	G
۱۱۶۲۸۴/۲	۱۱۶۶۹۷/۹	۱۱۶۶۷۰/۶	۱۱۶۶۹۰/۳	۱۱۶۹۸۶/۴	۱۱۶۹۰۴/۵	۱۱۶۷۰۹/۱	P
۱۱۶۳۱۴/۶	۱۱۶۷۲۲/۳۷	۱۱۶۶۹۵/۵۸	۱۱۶۷۱۴/۷۷	۱۱۷۰۱۲/۱۹	۱۱۶۹۲۹/۴	۱۱۶۷۳۴/۱۵	D
۳۴/۶۹	۳۴/۴۱	۳۴/۵۴	۳۴/۳	۳۴/۹۳	۳۴/۳	۳۴/۳۸	MSPE
۱۵۷۷/۹	۳۳۹/۰۲	۱۱۲/۵۳	۳۷۰/۴۶	۱۰۳/۷۵	۳۸۴/۳۱	۱۱۷/۴۷	زمان اجرا

بحث و نتیجه‌گیری

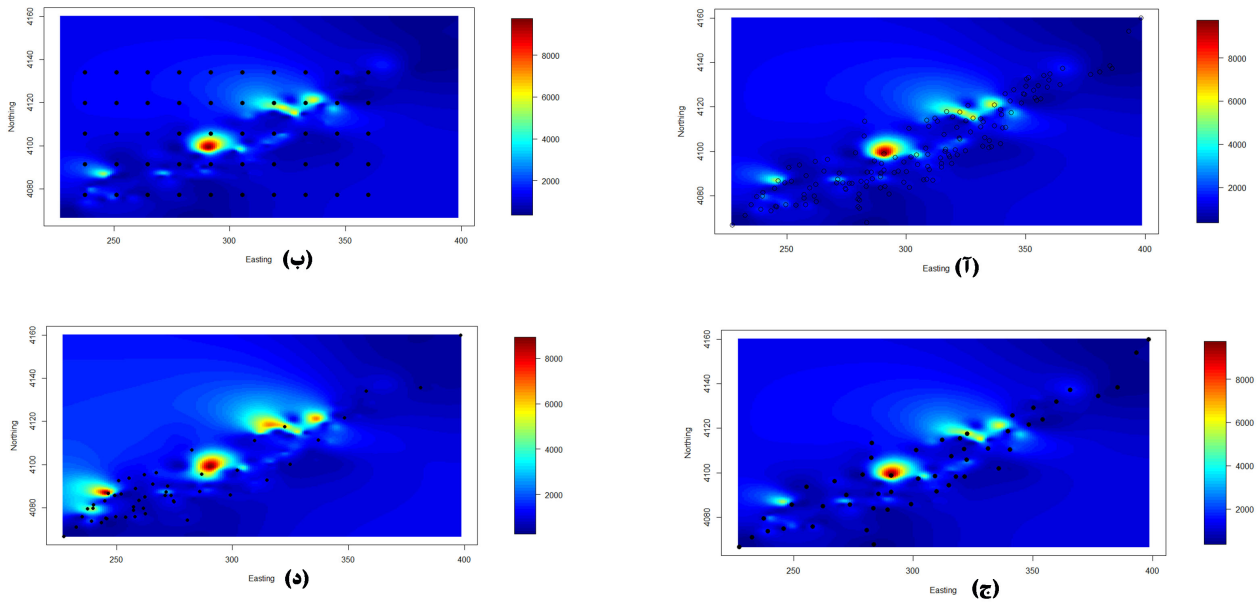
در این مقاله روش‌های مختلف انتخاب گره را معرفی کردیم و آنها را بر روی داده‌های کیفیت آب استان گلستان برای برازش مدل‌های فرایند پیشگو، پیاده کردیم.

با توجه به نتایج تحلیل داده‌ها می‌توان گفت که در عمل اگر داده‌ها به‌طور همگن در ناحیهٔ تحت مطالعه پراکنده باشند، تفاوت چندانی در روش‌های انتخاب گره وجود ندارد. در مقابل، تعداد گره‌ها (که در این مقاله ۲۵ و ۵۰ انتخاب شده بودند) تأثیر زیادی بر دقت برآورد پارامترها و پیشگویی‌های بعدی دارد؛ البته همچنین باعث افزایش هزینهٔ محاسباتی می‌شود. بنا بر این انتخاب تعداد نقاط گره باید با توجه به حساسیت دقت استنباط‌ها و هزینهٔ محاسباتی تعیین شود. پس، در عمل باید تحلیل را با تعداد گره‌های متفاوت انجام داد و نتایج را با هم، بر اساس معیارهای تعریف‌شده، مقایسه کرد. مقایسه باید نسبی باشد و زمان اجرای محاسبات در فرایند استنباط در نظر گرفته شود.

با مدل‌بندی داده‌های واقعی کیفیت آب‌های زیرزمینی استان گلستان، توانستیم عملکرد مدل‌های دون‌رتبه را به نمایش درآوریم که یک تقریب بسیار خوب از مدل کلّی می‌باشد. در واقع نتایج به‌دست‌آمده از مدل فرایند پیشگو تفاوت چشمگیری با مدل پر رتبه (مدل کلی و بدون کاهش رتبه) ندارد. این تقریب قابل قبول با کاهش هزینهٔ محاسباتی نیز همراه است. در واقع هزینهٔ محاسباتی از ۱۵۷۷/۹۰ ثانیه برای مدل پر رتبه به ۳۳۹/۰۲ ثانیه در بیشترین حالت برای مدل دون‌رتبه تقلیل یافته است و این زمان قابل توجهی را برای کاربر ذخیره می‌کند. بنا بر این نتایج نشان دادند که در کنار به دست آوردن کارایی محاسباتی قابل توجه، کارایی آماری از دست رفته در نتایج استنباط (با انتخاب مناسب تعداد گره‌ها) قابل صرف نظر کردن است. این نتیجه در مسائل کاربردی واقعی، که با داده‌های حجیم سر و کار دارند، می‌تواند بسیار مهم باشد. بنا بر این، در مجموعه‌داده‌هایی که ویژگی‌های آنها مشابه مواردی است که در این مقاله به آنها پرداخته شد، استفاده از چارچوب مدل‌های دون‌رتبه را به‌عنوان یک انتخاب توصیه می‌کنیم.

سپاسگزاری

از شرکت سهامی آب منطقه‌ای گلستان بابت در اختیار گذاشتن داده‌های آب زیرزمینی استان گلستان قدردانی می‌کنیم.



شکل ۱. مکان قرارگیری مشاهدات در سطح استان گلستان، همراه با رویهٔ پیشگویی‌شدهٔ میدان تصادفی با مدل رتبهٔ کامل (آ)؛ مکان قرارگیری گره‌های انتخابی با روش مشبکهٔ منظم با تعداد ۵۰ گره همراه با رویهٔ پیشگویی‌شده با مدل فرایند پیشگو (ب)؛ مکان قرارگیری گره‌های انتخابی با روش طرح پر کردن فضا با تعداد ۵۰ گره، همراه با رویهٔ پیشگویی‌شده میدان تصادفی با مدل فرایند پیشگو (ج)؛ مکان قرارگیری گره‌های انتخابی با روش تقریب طراحی بهینه با تعداد ۵۰ گره، همراه با رویهٔ پیشگویی‌شده میدان تصادفی با مدل پیشگو (د). برای راحتی و وضوح بیشتر در تصاویر، طول و عرض (بر حسب utm) تقسیم بر ۱۰۰۰ شده است.

مراجع

[1] Banerjee, S., Carlin, C.P. and Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall/CRC, Boca Raton, Florida.

[2] Banerjee, S., Gelfand, A.E., Finely, A.O. and Sang, H. (2008). Gaussian predictive process models for large spatial data sets, *Journal of the Royal Statistical Society, Series B*, **70**, 825-848.

[3] Barry, R.P. and Ver Hoef, J.M. (1996). Blackbox kriging: kriging without specifying variogram models, *Journal of Agricultural, Biological and Environmental Statistics*, **1**, 297–322.

[4] Berk, R.A. (2008). *Statistical Learning from a Regression Persepective*, Springer, New York.

[5] Chipman, H.A., George, E.I. and McCulloch, R.E. (2010). BART: Bayesian additive regression trees, *Annals of Applied Statistics*, **4**, 266–298.

[6] Crainiceanu, C.M., Diggle, P.J. and Rowlingson, B. (2008). Bivariate binomial spatial modeling of loa loa prevalence in tropical Africa (with discussion), *Journal of the American Statistical Association*, **103**, 21–37.

[7] Cressie, N. (1993). *Statistics for Spatial Data*, John Wiley, New York.

[8] Cressie, N. and Wikle, C. (2011). *Statistics for Spatio-Temporal Data*, John Wiley, London.

[9] Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large data sets, *Journal of the Royal Statistical Society, Series B*, **70**, 209–226.

[10] Diggle, P. and Lophaven, S. (2006). Bayesian geostatistical design, *Scandinavian Journal of Statistics*, **33**, 53–64.

[11] Finely, A.O., Sang, H., Banerjee, S. and Gelfand, A.E. (2009). Improving the performance of predictive process modeling for large datasets, *Computational Statistics and Data Analysis*, **53**, 2873-2884.

[12] Fuentes, M. (2007). Approximate likelihood for large irregularly spaced spatial data, *Journal of the American Statistical Association*, **102**, 321–331.

[13] Furrer, R., Genton, M.G. and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets, *Journal of Computational and Graphical Statistics*, **15**, 502–523.

[14] Gelfand, A.E. and Ghosh, S.K. (1998). Model choice: a minimum posterior predictive loss approach, *Biometrika*, **85**, 1–11.

[15] Gelfand, A.E., Kim, H., Sirmans, C.F. and Banerjee, S. (2003). Spatial modelling with spatially varying coefficient processes, *Journal of the American Statistical Association*, **98**, 387–396.

[16] Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2013). *Bayesian Data Analysis*, Third edition, Chapman & Hall/CRC, New York.

[17] Gu, C. (2002). *Smoothing Spline ANOVA Models*, Springer, New York.

[18] Higdon, D. (2002). Space and space-time modeling using process convolutions. In: Anderson, C., Barnett, C., Chatwin, P.C. and El-Shaarawi, A.H.: *Quantitative methods for current environmental issues*, Springer, Berlin, 37–56.

[19] Kammann, E.E. and Wand, M.P. (2003). Geoadditive models, *Journal of the Royal Statistical Society, Series C*, **52**, 1–18.

[20] Kaufman, C.G., Schervish, M.J. and Nychka, D.W. (2009). Covariance tapering for likelihood-based estimation in large spatial data sets, *Journal of the American Statistical Association*, **103**, 1545–1555.

[21] Lang, S. and Brezger, A. (2004). Bayesian P-splines, *Journal of Computational and Graphical Statistics*, **13**, 183–212.

[22] Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R. and Klein, B. (2000). Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV, *The Annals of Statistics*, **28**, 1570-1600.

[23] Nychka, D. and Saltzman, N. (1998). Design of air-quality monitoring networks, *In Case Studies in Environmental Statistics*, Lecture Notes in Statistics, eds D. Nychka, L. Cox and W. Piegorsch. Springer, New York, 51–76.

[24] Paciorek, C.J. and Schervish, M.J. (2006). Spatial modelling using a new class of nonstationary covariance functions, *Environmetrics*, **17**, 483–506.

[25] Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*, Second edition, Springer, New York.

[26] Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*, Second edition, Springer, New York.

[27] Royle, J. and Nychka, D. (1998). An algorithm for the construction of spatial coverage designs with implementation in S-PLUS, *Computers and Geosciences*, **24**, 479-488.

[28] Ruppert, D., Wand, M.P. and Caroll, R.J. (2003). *Semiparametric Regression*, Cambridge University Press, Cambridge.

[29] Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory of Kriging*, Springer Series in Statistics, New York.

[30] Stein, M.L., Chi, Z. and Welty, L.J. (2004). Approximating likelihoods for large spatial data sets, *Journal of the Royal Statistical Society*, Series B, **66**, 275–296.

[31] Stein, M.L. (2007). Spatial variation of total column ozone on a global scale, *Annals of Applied Statistics*, **1**, 191–210.

[32] Stein, M.L. (2008). A modeling approach for large spatial datasets, *Journal of the Korean Statistical Society*, **37**, 3–10.

[33] Vecchia, A. (1988). Estimation and model identification for continuous spatial processes, *Journal of the Royal Statistical Society*, Series B, **50**, 297–312.

[34] Wackernagel, H. (2006). *Multivariate Geostatistics: An Introduction with Applications*, Third edition, Springer, New York.

[35] Wikle, C.K. and Cressie, N. (1999). A dimension reduced approach to space-time Kalman filtering, *Biometrika*, **86**, 815–829.

[36] Xia, G. and Gelfand, A.E. (2006). Stationary process approximation for the analysis of large spatial datasets, Technical Report, ISDS, Duke University, Durham, NC.

[37] Xia, G., Miranda, M.L. and Gelfand, A.E. (2006). Approximately optimal spatial design approaches for environmental health data, *Environmetrics*, **17**, 363–385.

[38] Zhu, Z. and Stein, M. (2005). Spatial sampling design for parameter estimation of the covariance function, *Journal of Statistical Planning and Inference*, **134**, 583–603.